УДК 512.533.21

А. В. Клепинин

О СИНТАКСИЧЕСКИХ КОНГРУЭНЦИЯХ РАВНОМЕРНО РЕКУРРЕНТНЫХ ЯЗЫКОВ

Формальные языки являются важным объектом исследований современной математики. Это вызвано их активным использованием в самых разнообразных сферах: от абстрактных математических и прикладных вычислительных задач до исследований в таких, казалось бы, далеких от математики областях, как, например, биология или химия.

В данной работе рассматривается один из важных классов формальных языков – класс равномерно рекуррентных языков. Он интересен, в частности, тем, что в него попадают многие языки, получающиеся при помощи D0L систем (т. е. языки подслов последовательностей, порождаемых итерацией эндоморфизмов; см., например, [1]).

Каждому формальному языку соответствует алгебраический объект, характеризующий язык с точки зрения сложности его распознавания. Этот объект называется синтаксической конгруэнцией, и он играет важную роль в изучении и классификации формальных языков, поскольку фактически определяет алгоритм распознавания языка.

Основным результатом данной работы является описание синтаксических конгруэнций равномерно рекуррентных языков. В более слабой формулировке этот результат анонсировался на конференции «Дискретный анализ и исследование операций DAOR-02» (см. [2]).

Напомним необходимые понятия (точные формулировки см., например, в [3]). Зафиксируем конечный *алфавит* Σ. *Слово* над алфавитом Σ – это конечная цепочка, состоящая из букв алфавита Σ. Длину слова w будем обозначать через $|w|$. Через Σ^* (соответственно через Σ^+) обозначим свободную полугруппу (соответственно свободный моноид) над Σ относительно операции конкатенации. *(Формальным) языком* называется произвольное подмножество из Σ^*.

Пусть x и y – слова и для некоторых $x_1, x_2 \in \Sigma^*$ имеет место равенство $y = x_1 x x_2$. Тогда x называется *подсловом* y. В этом случае будем использовать обозначение $x \leq y$ (соответственно $x < y$, если $x \leq y$ и $x \neq y$).

Основное внимание в данной работе будет уделено конгруэнциям. Напомним общее определение этого понятия. *Конгруэнцией* полугруппы S называется отношение эквивалентности, устойчивое относительно полугрупповой

операции. Иными словами, отношение эквивалентности θ является конгруэнцией на S, если из того, что $(x, y) \in \theta$ и $(u, v) \in \theta$, следует, что $(xu, yv) \in \theta$ для любых элементов x, y, u и v полугруппы S. В частности, это означает, что на классах эквивалентности конгруэнции индуцируется полугрупповая структура (получается так называемая фактор-полугруппа по конгруэнции θ), а изучение исходной полугруппы оказывается сведено к изучению устройства классов конгруэнции и к изучению устройства фактор-полугруппы.

Язык L называется *равномерно рекуррентным*, если для произвольного слова $v \in L$ существует такое целое число n_v, что v содержится в качестве подслова во всяком слове $w \in L$ длины, большей n_v.

Заметим, что часто определение равномерной рекуррентности формулируется следующим образом. Язык L называется *равномерно рекуррентным*, если для любого числа n можно указать число $K(n)$ такое, что любое слово w длины $K(n)$ из языка L содержит все слова языка L длины, не большей n, или, что то же самое, если истинен следующий предикат:

$$\forall\, n \in \mathbb{N} \quad \exists\, K(n) \in \mathbb{N} \quad \forall\, w \in L \quad \forall\, v \in L \quad (|w| = K(n) \Rightarrow v \leq w).$$

Легко понять, что оба определения эквивалентны. Тем не менее в данной работе удобнее воспользоваться первым определением равномерной рекуррентности.

Язык называется *факториальным*, если он замкнут относительно взятия подслов. То есть если вместе с каждым словом $w \in L$ в языке L лежат все подслова $v \leq w$.

Обратим внимание на одну особенность факториальных языков. Легко проверяется, что если L – это факториальный язык над алфавитом Σ, то множество $I_L = \Sigma^+ \setminus L$ является идеалом в свободной полугруппе Σ^+.

Этому идеалу соответствует конгруэнция, которую мы будем обозначать через ρ_L. Одним из ее классов является идеал I_L (этот класс состоит из всех слов, не лежащих в L), а остальные классы одноэлементны (каждый такой одноэлементный класс состоит из какого-то одного слова языка L). То есть формальное определение ρ_L выглядит так: $(u, v) \in \rho_L$ тогда и только тогда, когда либо $u = v$, либо $u, v \in I_L$. Конгруэнция ρ_L – это *рисовская конгруэнция* полугруппы Σ^+ по идеалу I_L.

Для любого языка L (не обязательно факториального) можно определить *синтаксическую конгруэнцию* σ_L на Σ^+ следующим образом: $(u, v) \in \sigma_L$ тогда и только тогда, когда для произвольных $p, q \in \Sigma^*$ слова puq и pvq одновременно лежат или не лежат в L.

В случае факториальных равномерно рекуррентных языков эти два типа конгруэнций тесно связаны. Но прежде чем дать точную формулировку полученного результата, рассмотрим один частный случай.

Слово p называют *примитивным*, если его невозможно представить в виде степени какого-либо другого слова. Легко проверяется, что любое слово w можно представить в виде $w = p^n$ для подходящего примитивного слова p.

Пусть $p = a_1a_2 \dots a_k$ – некоторое примитивное слово. Рассмотрим язык L, устроенный следующим образом: $L = \cup_{n=1}^{\infty}\{w|midw \le p^n\}$. Будем называть языки такого типа *k-периодическими*. Непосредственно из определения следует, что язык L является факториальным и равномерно рекуррентным. Попытаемся понять, как устроена синтаксическая конгруэнция k-периодического языка.

Рассмотрим множество $C(k) = \{c_{ij} \mid 1 \le i, j \le k\} \cup \{0\}$. Определим на этом множестве умножение следующим образом:

$$c_{ij} \cdot c_{\ell m} = \begin{cases} c_{im}, & \ell = j+1 \\ 0, & \ell \ne j+1 \end{cases}, \qquad 0 \cdot c_{ij} = c_{ij} \cdot 0 = 0 \cdot 0 = 0.$$

Непосредственно проверяется, что относительно введенной операции $C(k)$ является полугруппой.

Определим отображение $\varphi_L \colon \Sigma^+ \to C(k)$ следующим образом. Если непустое слово w лежит в k-периодическом языке L, то оно представляется в виде: $w = a_ia_{i+1} \dots a_kp^na_1a_2 \dots a_j$. Положим $\varphi_L(w) = c_{ij}$. Для всех $v \notin L$ положим $\varphi_L(v) = 0$. Непосредственно проверяется, что φ_L – гомоморфизм полугрупп, синтаксическая конгруэнция σ_L является ядром этого гомоморфизма, а синтаксическая полугруппа Σ^+/σ_L изоморфна $C(k)$.

Теперь мы можем сформулировать основной результат.

Теорема 1. *Если L – бесконечный факториальный равномерно рекуррентный язык, то либо его синтаксическая конгруэнция совпадает с рисовской по идеалу $\Sigma^+ \setminus L$, либо L является k-периодическим языком и его синтаксическая полугруппа изоморфна $C(k)$, а синтаксическая конгруэнция совпадает с ядром гомоморфизма φ_L.*

Для доказательства теоремы нам потребуется ряд промежуточных рассуждений.

Лемма 1. *Пусть L – факториальный язык. Тогда $\rho_L \subseteq \sigma_L$.*

Доказательство. Непосредственно из определения конгруэнции ρ_L следует, что язык L является объединением ее классов. И поскольку известно (см., например, [4]), что синтаксическая конгруэнция языка является наибольшей конгруэнцией с таким свойством, то $\rho_L \subseteq \sigma_L$.

Лемма 2. *Конгруэнции σ_L и ρ_L факториального языка L совпадают тогда и только тогда, когда все классы конгруэнции σ_L, составляющие язык L, одноэлементны.*

Доказательство. В одну сторону утверждение доказывается тривиально, поскольку из равенства конгруэнций $\sigma_L = \rho_L$ и определения конгруэнции ρ_L немедленно следует одноэлементность классов σ_L, составляющих L.

Обратно. Вложение $\rho_L \subseteq \sigma_L$ следует из леммы 1. Покажем, что и вложение $\rho_L \supseteq \sigma_L$ имеет место.

Пусть $(u, v) \in \sigma_L$. Если $u \notin L$, то из определения синтаксической конгруэнции $v \notin L$. Но это значит, что $u, v \in I_L$ и $(u, v) \in \rho_L$. Если же $u \in L$, то из предположения, что все классы конгруэнции, составляющие язык L, одноэлементны, следует, что $u = v$ и $(u, v) = (u, u) \in \rho_L$.

Лемма 3. *Пусть L – язык, $u, v \in L$, $v = puq$, $(u, v) \in \sigma_L$ и $w_n = p^n u q^n$. Тогда для всех $n \in \mathbb{N}_0$ слова w_n лежат в L. Более того, все они лежат в одном классе конгруэнции σ_L.*

Доказательство. Заметим, что для произвольного n справедливы равенства:

$$w_{n-1} = p^{n-1}\, u\, q^{n-1},$$
$$w_n = p^{n-1}\, puq\, q^{n-1} = p^{n-1}\, v\, q^{n-1}.$$

Поскольку $(u, v) \in \sigma_L$, то $(w_{n-1}, w_n) \in \sigma_L$ и, как следствие, $(w_n, u) \in \sigma_L$. Но слово u лежит в языке L. Следовательно, и весь σ_L-класс слова u, содержащий все w_n, лежит в L.

Лемма 4. *Если L – факториальный язык, $u, v \in L$, $u < v$, $(u, v) \in \sigma_L$, то язык L содержит все степени некоторого слова $x \in L$.*

Доказательство. Поскольку $u < v$, положим $v = puq$. Отметим, что слова p и q не могут быть одновременно пустыми, так как $u \neq v$. Без ограничения общности пусть p не является пустым. Положим $x = p$. Тогда из леммы 3 следует, что для всех натуральных n слова $x^n u q^n$ лежат в L. И в силу факториальности L это означает, что $x^n \in L$ для всех n.

Для доказательства следующей леммы нам потребуется понятие лексикографического порядка. Если на Σ задан некоторый линейный порядок (например, $\Sigma = \{\alpha_1, \alpha_2, \ldots, \alpha_t\}$ и $\alpha_1 < \alpha_2 < \cdots < \alpha_t$), то на Σ^+ индуцируется *лексикографический порядок* $<_\ell$, определяемый следующим образом. Пусть $v, w \in \Sigma^+$ и $v \neq w$. Если $v = u\alpha_i v'$, $w = u\alpha_j w'$ и $\alpha_i \neq \alpha_j$, то по определению полагаем $v <_\ell w$ тогда и только тогда, когда $\alpha_i < \alpha_j$. Иначе, $v <_\ell w$ тогда и только тогда, когда $|v| < |w|$.

Лемма 5. *Если язык L бесконечен и равномерно рекуррентен, то в одном классе его синтаксической конгруэнции не может быть двух слов равной длины.*

Доказательство. Предположим обратное: пусть $(x, y) \in \sigma_L$, $x, y \in L$, $x \neq y$, $|x| = |y|$.

Зафиксируем на Σ некоторый линейный порядок. Тогда на Σ^+ имеем лексикографический порядок $<_\ell$. Без ограничения общности можно считать, что $x <_\ell y$ (т. е. $x = u\alpha x'$, $y = u\beta y'$, $\alpha, \beta \in \Sigma$ и $\alpha < \beta$).

Воспользуемся равномерной рекуррентностью языка L и зафиксируем n такое, что в любом слове $w \in L$ длины, не меньшей n, встречается подслово x.

Поскольку язык L бесконечен, в нем найдется слово w_0 такое, что $|w_0| \geq n$. Тогда $w_0 = p_0 x q_0$.

Положим $w_1 = p_0 y q_0$. Тогда поскольку $(x, y) \in \sigma_L$, то и $(w_0, w_1) \in \sigma_L$. Следовательно, $w_1 \in L$.

Заметим, что длины w_0 и w_1 совпадают, но при этом w_0 лексикографически меньше w_1.

Поскольку $|w_1| \geq n$, в слове w_1 присутствует x в качестве подслова, т. е. $w_1 = p_1 x q_1$. Положим $w_2 = p_1 y q_1$. Тогда опять из $(x, y) \in \sigma_L$ следует, что $(w_1, w_2) \in \sigma_L$, $w_2 \in L$, и вновь $|w_2| = |w_1| \geq n$.

Повторяя рассуждения, получим цепочку слов $\{w_k\}_{k \in \mathbb{N}_0}$ такую, что выполняются свойства:

1) $|w_{k+1}| = |w_k| = |w_0|$;

2) w_{k+1} лексикографически больше, чем w_k.

Но бесконечная цепочка с указанными свойствами существовать не может, поскольку в конечном множестве существуют лишь конечные строго возрастающие цепи. Противоречие.

Лемма 6. *Если L – бесконечный равномерно рекуррентный язык, то либо классы его синтаксической конгруэнции, объединением которых он является, одноэлементны, либо найдутся такие слова $s, t \in L$, что $s < t$ и $(s, t) \in \sigma_L$.*

Доказательство. Допустим, что среди классов синтакической конгруэнции, составляющих L, есть неодноэлементный класс, т. е. $(x, y) \in \sigma_L$, $x, y \in L$, $x \neq y$.

Из леммы 5 следует, что $|x| \neq |y|$. Без ограничения общности будем считать, что $|x| < |y|$. Воспользуемся равномерной рекуррентностью языка L и

зафиксируем n такое, что в любом слове $w \in L$ длины, не меньшей n, встречается подслово x.

Поскольку язык L бесконечен, в нем найдется слово w_0 такое, что $|w_0| \geq n$. Тогда $w_0 = p_0 x q_0$. Положим $w_1 = p_0 y q_0$. Тогда поскольку $(x, y) \in \sigma_L$, то и $(w_0, w_1) \in \sigma_L$. Следовательно, $w_1 \in L$.

Заметим, что $|w_1| = |p_0| + |y| + |q_0| > |p_0| + |x| + |q_0| = |w_0| \geq n$. Это означает, что в слове w_1 присутствует x в качестве подслова, т. е. $w_1 = p_1 x q_1$. Положим $w_2 = p_1 y q_1$. Тогда опять из $(x, y) \in \sigma_L$ следует, что $(w_1, w_2) \in \sigma_L$, $w_2 \in L$. И опять $|w_2| > |w_1| \geq n$.

Повторяя рассуждения, получим цепочку слов $\{w_k\}_{k \in \mathbb{N}_0}$ такую, что выполняются свойства:

1) $|w_{k+1}| > |w_k|$;

2) все w_k лежат в одном классе конгруэнции σ_L;

3) все w_k являются словами из L.

Поскольку язык L является равномерно рекуррентным, из возрастания $|w_k|$ следует, что найдется такое n, что $w_0 \leq w_n$. Положим $s = w_0$ и $t = w_n$. Тогда $s, t \in L$, $s < t$ и $(s, t) \in \sigma_L$.

Теперь все готово для доказательства основного результата.

Доказательство. В силу леммы 2 совпадение конгруэнций σ_L и rho_L равносильно тому, что все классы конгруэнции σ_L, составляющие язык L, одноэлементны. Поэтому если совпадения конгруэнций нет, то в силу леммы 6 найдутся такие слова $s, t \in L$, что $s < t$ и $(s, t) \in \sigma_L$.

Тогда из леммы 4 немедленно следует, что язык L содержит все степени некоторого слова $p \in L$. Без ограничения общности можно считать, что слово p – примитивное.

В силу факториальности L имеем включение $L \supseteq \cup_{n=1}^{\infty} \{w \mid w \leq p^n\}$.

Теперь пусть $w \in L$. Поскольку L – равномерно рекуррентный, зафиксируем k такое, что в любом слове $v \in L$ длины, не меньшей k, встречается подслово w. Но тогда немедленно получаем, что $w \leq p^k$. Следовательно, мы получили обратное включение: $L \subseteq \cup_{n=1}^{\infty} \{w \mid w \leq p^n\}$. Таким образом, L – $|p|$-периодический язык, что, с учетом замечаний, сделанных перед теоремой, завершает доказательство.

В заключение отметим, что вопрос совпадения (или несовпадения) синтаксической конгруэнции языка с рисовской конгруэнцией по его дополнению

уже изучался в ряде работ. Так, например, ранее свойство совпадения синтаксической конгруэнции языка с рисовской конгруэнцией по дополнению до языка было выявлено у двухбуквенного языка Туэ-Морса (см. [5]), а затем у языка Аршона (см. [6]). Теорема, доказанная в данной работе, обобщает эти результаты и выявляет причины, по которым названные языки обладают свойством совпадения названных конгруэнций.

Работу [5] следует выделить особо, поскольку в ней получено несколько достаточных условий несовпадения названных конгруэнций. Эти условия, в частности, позволяют легко сконструировать примеры, демонстрирующие, что названные конгруэнции могут не совпадать. Что более важно, в работе [5] аргументируется необходимость исследования названных конгруэнций, поскольку результаты в данной области позволяют продвинуться в описании свойств так называемых 0-приведенных многообразий полугрупп. А объединение результатов настоящей работы и [5] фактически добавляет к набору достаточных условий несовпадения названных конгруэнций необходимое условие несовпадения.

Также следует отметить, что совпадение синтаксической конгруэнции некоторого языка с рисовской конгруэнцией фактически означает, что с точки зрения распознавания язык устроен достаточно сложно и что переход к изучению синтаксической полугруппы и классов синтаксической конгруэнции не позволяет упростить задачу. Следовательно, для построения алгоритмов распознавания подобных языков следует искать иные подходы.

Литература

1. Lothair M. Algebraic Combinatorics on Words. Cambridge: Cambridge Univ. Press, 2002.
2. Клепинин А. В. О синтаксических конгруэнциях бесповторных языков над конечными алфавитами // Дискретный анализ и исследование операций: Материалы Росс. конф. Новосибирск, 2002. С. 129.
3. Choffrut C., Karhumäki J. Combinatorics on words // Handbook of Formal Languages. Berlin, 1997. Vol. 1. P. 329–438.
4. Лаллеман Ж. Полугруппы и комбинаторные приложения. М.: Мир, 1985.
5. Суханов Е. В., Шур А. М. Об одном классе формальных языков // Алгебра и логика. 1998. Т. 37, № 4. С. 478–492.
6. Клепинин А. В., Суханов Е. В. О комбинаторных свойствах языка Аршона // Дискретный анализ и исследование операций. 1999. Т. 6, № 2. С. 23–40.